Supplementary Information for
The Ancestry and Affiliations of Kennewick Man

Table of contents

1. Sample processing

Morten Rasmussen

1.1 Ancient sample, Kennewick Man

We received a small metacarpal bone fragment (see section 10 for sample specifics) from the Burke Museum of approximately 200mg, after lightly cleaning the surface with a Dremel drill, we powdered the sample for DNA extraction following published protocols[1,2]. Briefly, we dissolved the bone powder in digestion buffer (0.47M EDTA, with 0.5% N-laurylsarcosyl and proteinase K) by incubating 24hrs at 37C. The digest was then mixed with binding buffer (5M GuSCN, 0.05M Tris-HCl pH=8, 0.05M NaCl, 0.02M EDTA, 1% TritonX-100) and fine-grain silica, pH was adjusted to 4.0-5.0 using concentrated HCl. After 3 hrs of incubation silica was pelleted, washed in cold 80% ethanol and eluted into 100ul of EB (Qiagen, Germany)

4 libraries were prepared from the extract, 2 using a double strand library (DSL) method and 2 using the single strand library (SSL) method. For DSL a commercial kit from New England Biolabs (E6070, Ipswich, MA) was used with a modified protocol as published previously[3]. For SSL preparation we followed the protocol as published in Meyer et al.,[4] without USER enzyme treatment.

All libraries were amplified with Kapa HiFi Uracil+ (Kapa Biosystems, Woburn, MA) following the same cycling conditions as Rasmussen et al[3]. For SSL we used a shortened forward primer (SSL_forward 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTC-CCTACACGACGCTCTTCC), to match the ligated adaptor.

1.2 Modern samples, Colville Tribe

The saliva samples from the Colville, was collected by the tribe members themselves, after we supplied them with Oragene-DNA saliva sampling kits (OG-500, DNA Genotek, Ottawa, Canada). DNA was extracted from the saliva using prepIT-L2P extraction kit following manufacturers guidelines (DNA Genotek, Ottawa, Canada). The purified DNA was genotyped on HumanOmni5Exome BeadChips from Illumina (San Diego, CA); genotyping was performed at Aros Applied Biotechnology (Aarhus, Denmark). Written approval was collected from the tribe in addition to individual signed consent to participate in the study.

2. Initial data processing

Morten Rasmussen

A total of 47 lanes of HiSeq2000 sequencing were performed on the 4 libraries. With much higher complexity in the SSL the sequencing was split as 4 lanes for the DSL and 45 lanes for the SSL. Majority of the sequencing was run as single end runs with either 94 or 101 cycles, except 2 lanes of paired end 101 cycles for the SSL. In total we generated 6,020,538,382 reads for all libraries combined. Basecalling and demultiplexing were done using CASAVA-1.8.2

2.1 Trimming and mapping

Raw reads were processed with AdaptorRemoval-1.5.2[5] to remove sequenced adaptor, leading/trailing Ns and low quality bases. For the paired end data, reads were merged also using AdaptorRemoval, in all cases only reads of a minimum of 30bp

was kept. Trimmed reads were mapped to the human reference genome (Build 37.1) using bwa-0.7.5a-r405[6] with mapping seed disabled. Low quality mappings and duplicate reads were removed using samtools-0.1.19-44428cd[7], only keeping unambiguously mapped reads with a MAPQ>=30. Readgroups were added using PicardTools-1.108 (http://broadinstitute.github.io/picard/) before realigning with GATK- 2.8-1 and updating md-tag with samtools calmd, summary for each library can be found in table S1.

2.2 DNA Damage

It has been shown that over time cytosine will deaminate to uracil leading to a characteristic damage signal in ancient DNA samples[8], which is often used as a test for authenticity. To test for this pattern, and rescale quality scores, we ran mapDamage2[9]. As expected we get different results for the two types of library[4], where DSL have C→T on the 5' end and G→A on the 3' end (Fig. S1a), while SSL shows C→T on both ends (Fig. S1b).

2.3 mtDNA

To determine mitochondrial haplogroup of Kennewick Man, all trimmed reads were mapped against the revised Cambridge reference sequence using following the pipeline described above. A vcf-file with genotype calls was generated using samtools and bcftools-0.1.19-44428cd (http://samtools.sourceforge.net), only high quality calls with a quality score of 50 or more were kept. Mitochondrial haplogroup was called using HaploGrep 2 beta[10], followed by manual verification of each diagnostic variant. Kennewick Man is placed at the root of X2a, and does not share any of the derived alleles on the branches leading to X2a1 or X2a2. A phylogenetic tree (Extended Data Fig. 1) was constructed using a median-joining network as implemented in Network version 4.6.1.0[11], and subsequently refined manually. Nucleotide numbering is consistent with rCRS[12], with topology of the tree matching phylotree (build 16)[13], and similarly we excluded The mutations 309.1C(C), 315.1C, AC indels at 515-522, 16182C, 16183C, 16193.1C(C) and 16519 from the tree. Available sequences from haplogroup X were downloaded from GenBank, and a subset of the tree is visualized in Extended Data Fig 1.

An estimate of contamination on the mtDNA was generated using contamMix_1.0-10[14], with a 95% confidence interval ranging from 3.7-7.1%, somewhat higher than the X-chromosome based test. Except for mitochondrial haplogroup call, all tests are based on nuclear DNA sequences.

3. Molecular decay of the Kennewick Man

Morten E. Allentoft

The DNA sequence length distribution obtained from shotgun sequencing data can be used to investigate the molecular preservation in an ancient specimen[15]. In an aDNA extract there should be a negative exponential correlation between the number of DNA molecules and their length. This is an effect of fragmentation of the DNA strands, leaving few long DNA fragments and many short ones[15,16]. Following previous studies[15,17,18], we investigated only the declining part of the sequence length distribution to remove molecular artifacts caused by a poor recovery of short fragments in the DNA extraction, and a fixed maximum sequencing length (Figure S2). For comparative purposes the decay estimates were conducted first and foremost

based on data from double stranded DNA libraries (DSL) and the values discussed below refer to those data except where noted.

The Kennewick data conformed well to an exponential decay model ($R^2 = 0.94$) but the correlation was partly obscured by a 10 bp periodicity in the distribution (Figure S2). This phenomenon has been described previously in genomic data and is likely mirroring the 10 bp turn of the DNA helix combined with preferential strand cleavage of the DNA backbone facing away from nucleosome protection[19]. Deagle et al.[16] showed that the decay constant (λ) in the exponential relationship represents the DNA damage fraction. We estimated λ in the Kennewick genome to 0.017 (Figure S2b), implying that 1.7% of the phosphodiester bonds in the DNA backbone are broken. Moreover, $1/\lambda$ is equivalent to the expected mean DNA fragment length in the sample[16] and we calculated this to 59 bp (Table S2). We note that this is not the same as the mean sequence length in data which is biased both experimentally and bioinformatically.

It has been shown that long-term post mortem DNA fragmentation can be described as a rate process, and that the damage fraction (λ, per site) can be converted to a decay rate (k, per site per year), when the age of the sample is known[15]. Here we assume an age of 9,000 years (calibrated radiocarbon age, 9,075-8,935 CAL BP), yielding a rate of decay of 1.89E-6 per site per year, and a molecular half-life[15] of 3,670 years for 100 bp DNA fragments (Table S2). After this time, half the fragments of this length will have had one or more strand breaks. We compared this rate of molecular decay to that of other ancient genomes that have been characterized recently by our lab (Table S2). It is clear that the Kennewick Man genome falls within the range of previous results, with a slower molecular decay than observed in ancient human samples from Spain[17] and the Caribbean[20] but faster than observed in the Anzick-1 skeleton from Montana[3] (Table S2). Post mortem depurination of the DNA is a temperature-sensitive reaction that results in strand breakage[15,21-23], which is why ancient samples from warm climates will often yield none or only very little authentic identifiable DNA. This is clearly exemplified in Table S2, where samples from the Caribbean display a much faster decay rate than observed in the other samples. For example, the DNA decay rate in the Caribbean samples are c. 22 times faster than observed in the DNA from Kennewick Man.

Kennewick Man and Anzick-1 were from a comparable latitude, but the Anzick site in Montana is nonetheless considerably colder (Table S2), which is mirrored in our data by a slower rate of molecular decay (Table S2). The La Braña skeleton however, was preserved exposed on the surface in a cave in Spain, and although being c. 1,500 years younger and presumably preserved colder than Kennewick Man, the DNA is still more degraded with an average estimated fragment length of 30 bp compared to 59 bp in Kennewick. This shows that mean surface temperature and age can only account for some of the variation in DNA preservation.

We also examined the fragment length distribution based on single stranded libraries (SSL). As expected for ancient DNA, this is skewed towards shorter fragments compared to DSL data (Figure S2) because nicks in the DNA backbone will translate into shorter fragments when the DNA is denatured in the SSL preparation[4]. For the SSL data we had sufficient data to separate nuclear DNA from mtDNA in the

analyses (Figure S2). As described previously, the mtDNA shows better preservation than the nuclear DNA (Table S2). This is likely owing to the circular structure of the mitochondrial genome which reduces attack from exonucleases, and/or some degree of protection behind the double membrane of the mitochondrion[15].

In summary, we conclude that the molecular characteristics of the Kennewick Man sample correspond to the patterns expected for ancient degraded DNA, and serve as a general confirmation of the authenticity of this genome.

4. Estimating contamination and inferring the source population using X-chromosome data

Anders Albrechtsen and Ida Moltke

4.1 Method

To estimate the amount of contamination in the Kennewick Man sample and to infer the source population of the contamination, we used a model similar to the one described in Rasmussen et al[24]. The idea behind this model is to exploit the fact that Kennewick man is male and therefore is haploid in the X-chromosome (except in the pseudo-autosomal region). As a consequence, any discordance in observed bases in a X-chromosomal site observed in the read data from the Kennewick Man sample must be caused either by a sequencing error or contamination. We exploit this by looking at a fixed set of L sites on the X-chromosome that are known to be polymorphic. We assume that regardless of the source population the probability that the sites that are adjacent to these known polymorphic sites are also polymorphic is very low. We also assume that the error rate for the set of known polymorphic sites is the same as the error rate for the adjacent sites and note that if these assumptions hold true, then the base discordance rate for the known polymorphic sites should be the same as for the adjacent sites if there is no contamination. In contrast, contamination from any human source should lead to a higher discordance rate for the known polymorphic sites, whereas it should have little, if any, effect on the discordance rate in the adjacent sites. Hence by comparing discordance rate in the known polymorphic sites to the discordance rates in their adjacent sites, we can estimate the amount of potential contamination, essentially by quantifying how much higher the discordance rate is in the known polymorphic sites. We do this using a probabilistic model that takes the allele frequencies in the different possible source populations into account. Furthermore, by integrating this model into a Bayesian framework, we estimate posterior probabilities for different populations being the source population and based on this we infer the most probable source population.

4.2 Model

First, for any given site we define *major reads* as reads that at the site carry the base, which is most frequently observed at the site, and we define *minor reads* as reads that at the site carry any of the other 3 bases. Then, for any given known polymorphic site l on the X-chromosome, we let f_l denote the frequency of the major read at l and model the probability of observing the data $X_l = (N_M, N_m)$ consisting of the number of major reads N_M and minor reads N_m at site l using a binomial distribution $P(k; n, p)$:

$$P(X_l|\theta, \varepsilon, f_l) = P(k = N_m; n = N_m + N_M, p) \propto (1-p)^{N_M} p^{N_m}$$

were p is the probability of observing a minor read and depends on the error rate, ε, which we estimate based on the sites adjacent to l, and the contamination rate, θ. Assuming independence between sites we can, based on this, write the likelihood for L polymorphic sites on the X-chromosome as:

$$L(\theta, \varepsilon | f) = \prod_{l=1}^{L} P(X_l | \theta, \varepsilon, f_l)$$

where f is the vector of all site-specific frequencies f_l.

To model the probability of observing a minor read, $p = P(Z = 1|\theta, \varepsilon, f)$, we introduce the variable Z, which for a given read takes the value 1 if the read is a minor read and 0 otherwise. Additionally, we make the assumption that that both the contamination rate and error rate is low. This means that we can assume that the major read is the sample's true genotype and that the minor reads are either caused by error or contamination (variables E and C, respectively). It also means that we can assume that a base cannot be both a contaminant, $C = 1$, and an error $E = 1$. Consequently, we can write

$$p = P(Z = 1|\theta, \epsilon, f) = \sum_{c \in \{0,1\}} \sum_{e \in \{0,1\}} P(Z = 1|C = c, E = e, f) P(C = c|\theta) P(E = e\,|\varepsilon)$$

where

$$\mathrm{P(Z = 1|C = c, E = e, f)} = \begin{cases} 0 & if\ C = E = 0\ \vee\ C = E = 1 \\ 1 & if\ C = 0\ \wedge E = 1 \\ 1 - f & if\ C = 1\ \wedge E = 0 \end{cases}$$

and

$$\mathrm{P(C = c|\theta)} = \begin{cases} \theta, & if\ C = 1 \\ 1 - \theta, & if\ C = 0 \end{cases}$$

It should be noted that this model ignores that the contamination is likely from a single individual only and therefore the model is only valid if you only have less than two contamination reads per site. However, if the contamination rate is low it is very unlikely that there is more than one contamination read per site for low or medium sequencing depth.

4.2.1 Estimation of θ and inference of the source population

Based on the model described above we performed maximum likelihood estimation of θ with f set to the allele frequencies estimated from the assumed source population and ε set to the discordance rate in the sites adjacent to the known set of polymorphic sites (four sites on each side). Standard errors for the θ estimates were estimated using a jackknife procedure[25].

Furthermore, we used the described model to infer which population in a set of all possible source populations, *Popset,* is the source of the contamination. We did this by assuming a uniform prior on *Popset*, i.e. we assumed that all the populations in *Popset* are a priori equally likely to be the source population. Under this assumption, the posterior probability that a population y is the source population is according to Bayes rule given as:

$$P(pop = y|\theta, \varepsilon, f) = \frac{L(\theta, \varepsilon | f = f^y)}{\sum_{y' \in Popset} L(\theta, \varepsilon | f = f^{y'})}$$

where f^y is a vector of the allele frequencies in population y for all analyzed L polymorphic sites.

4.3 Data

We first applied the method to simulated data to ensure that it gives reasonable results. Then we applied it to the Kennewick Man sample. Below is a description of the data used for these analyses.

4.3.1 Allele frequencies and choice of *Popset*

For the set of possible source populations, *Popset,* we used the 1000 genomes data from five selected populations: French (CEU), Han Chinese (CHB), Indian (GIH), Peruvians (PEL) and Nigerians (YRI). Because some of the individuals in this dataset are admixed (see Figure S3) the allele frequencies used in our analyses were ancestral allele frequencies estimated using ADMIXTURE[26]. Only sites with a minor allele frequency above 5% and a missingness below 5% across all five populations were used and sites closer than 10 bases to another polymorphic site were removed.

4.3.2 Data used for simulated contamination

For the simulations two individuals sequenced by Meyer et al.[4] were used. Namely the European individual HGDP00521 (CEU) and the Native American individual HGDP00998 (Karitiana). A minimum base quality score of 20 and a minimum mapping quality of 30 was required for inclusion of data in the analyses. In addition, a minimum mappability (100mer) score of 1 was required and the 2.5Mb ends of the X-chromosome were discarded to remove the pseudo-autosomal region.

4.3.3 Kennewick Man sample

The Kennewick Man sample was filtered in the same way as the genomes used for the simulations of contamination. In total this left us with data for 6,909 polymorphic sites, which we had allele frequency estimates for and at least two reads for the Kennewick Man sample.

4.4 Simulations

To test the method we simulated data with varying degrees of contamination. A Karitiana and a European (CEU) individual were used. To simulate contamination varying amount of data in the Karitiana sample was replaced with data from the CEU sample. This was done by first subsampling observed bases for the Karitiana such that each observed base has a chance π of being removed. Then a subsample of the CEU individual was added by sampling a base with probability $\pi N_{Karitiana}/N_{CEU}$, where N is the total number of bases present in each individual. The results of applying the method to the simulated data are shown in Figure S4. Note that without adding contamination the Karitiana sample appears to have 0.02% contamination from an African source. And importantly the method performs well.

4.5 Results for the Kennewick Man sample

The Kennewick Man sample appears to be contaminated. As can be seen in Table S3 there are markedly more minor reads at polymorphic positions than at the neighboring positions. And from Table S4 it can be seen that the contamination appears to be European; the posterior probability for this is 0.9565 as opposed to at most 0.025 for the other populations considered. The estimated contamination rate assuming CEU as the contamination source population is 2.5% (s.e. 0.44) (Table S4).

5. Y-Chromosome Lineage

G. David Poznik

Upon probing the sequence data for phylogenetically informative SNPs from the ISOGG database (http://www.isogg.org), we established that the Kennewick Man possessed Y-chromosome haplogroup Q (hgQ), the preponderant haplogroup among Native Americans[27,28].

To gain greater resolution, we used the Y-chromosome phylogeny constructed for the study of a Late Pleistocene human from a Clovis burial site[3]. In addition to the ancient sequence (Anzick-1), the tree included sequences from Human Genome Diversity Panel samples[29] sequenced in Poznik et al.[30], an ancient Saqqaq Palaeo-Eskimo[24], and the 11 hgQ samples from phase 1 of the 1000 Genomes Project[31] (Extended Data Fig. 1a).

We observed exclusively derived alleles at the 30 haplogroup-P SNPs for which Kennewick read data existed (Extended Data Fig. 1b). This branch (#30) is immediately ancestral to haplogroup Q. Of the 33 genotypes observed at SNPs between the origin of hgQ and the emergence of subgroup Q-M3 (branches 28 and 26), 32 were in the derived state. Read data were available for 3 of 17 SNPs on the Q-M3 branch, and all were derived, thereby definitively establishing the Kennewick Man as a member of hgQ-M3, a lineage observed exclusively among Native Americans[32] and in Northeast Siberia[33].

We observed exclusively ancestral alleles at all sites on the remaining branches of the tree, as well as at all sites within the Q-M3 subtree of the phase 3 1000 Genomes sample, which includes an additional 23 individuals.

6. Datasets and Masking

José Victor Moreno Mayar

6.1 Native American genotype dataset

We assembled a Native American genotype dataset consisting of 577 individuals from 59 different ethnic groups, genotyped over 300,934 sites. The dataset is an intersection of 490 individuals[34] (52 ethnic groups), 66 individuals[35] (6 ethnic groups), as well as 21 individuals from the Colville tribe that we genotyped for this study. Given some individuals in the panel have varying degrees of European and African admixture, we masked regions in their genomes where they are not homozygous for Native American ancestry. This allowed us to focus our analysis only on the Native American component of the data and to avoid any confounding signal that non-Native American admixture may produce.

First, we merged the Native American panel with a genotype reference panel representative of European, African and Native American ancestry. The reference panel consisted of (1) 30 CEU individuals[36] , (2) 30 YRI individuals[36] and (3) 30 Native American individuals[34]. The 30 Native American reference individuals were randomly chosen from a subset with no detectable European or African ancestry from an *ADMIXTURE*[26] run with K=3. We then phased the merged dataset using *shapeit2*[37]. We used the 1000 genomes phased variant panel (Phase I v3) as a reference and the HapMap recombination rates as a proxy for the human genetic map. We then inferred local ancestry from the three ancestral populations (Europeans,

Africans and Native Americans) using *RFmix*[38] with the G parameter set to 15 generations and allowing for phase correction. Finally, for each individual, we masked every region containing at least one European or African allele, according to the *RFmix* Viterbi calls.

6.2 Native American + Ainu + Polynesian + Siberian + outgroup genotype dataset

In order to have a dataset suitable for determining the genetic ancestry of the Kennewick man, we merged our Native American panel with genotype data from populations that have been hypothesized to be the source population for such remains; namely the Ainu and Polynesian islanders. Our merged dataset includes the 577 masked Native American individuals, 36 Ainu individuals[39], 33 Polynesian islanders[40,41] where we could not detect European ancestry, as well as 108 CEU, 83 CHB and 109 YRI individuals[3] that serve as outgroups. We furthermore included data of 161 Siberian individuals from 17 populations[24,42], excluding individuals with recent European admixture. The final merged dataset included a total of 1,107 individuals, genotyped over 62,923 sites.

7. Population genetic analyses

Martin Sikora

7.1 Kennewick Man and Anzick-1 genotyping

Genotypes for Kennewick Man and Anzick-1 were obtained from aligned reads at all SNP positions in the reference datasets. For the low-coverage Kennewick Man data, we randomly sampled a single read with both mapping and base quality ≥ 30. For the higher coverage Anzick-1 data[3], we called diploid genotypes using the 'call' command of bcftools (https://github.com/samtools/bcftools) and filtering for quality score (QUAL) and genotype quality (GQ) ≥ 30. Any variant where alleles for the ancient individuals did not match either of the alleles observed in the reference panel were discarded. To investigate the effects of the lower coverage in the Kennewick Man data on the ancestry analyses, we generated a randomly subsampled bam file from the high coverage Anzick-1 genome[3] approximately matching the coverage of Kennewick Man (1.6X at reference panel SNP positions), using the 'view' command of samtools with the '-s' flag (-s 0.05). Genotypes were then obtained by randomly sampling a single read as described above.

7.2 Principal component analysis

Principal component analysis was performed on a subset of individuals excluding the 109 YRI individuals, using EIGENSOFT[43]. The two ancient individuals were projected onto the components inferred from these sets of modern individuals by using the 'lsqproject' option of *smartpca*. Heterozygote genotypes were converted to homozygous prior to the analysis, by randomly sampling a single allele at each locus and individual. The PCA results for the subsampled Anzick-1 genome (Figure S5a) demonstrate that the effects of the lower coverage of the Kennewick Man are negligible, with only a slight shift towards European populations observed. Under the assumption of a higher fraction of contaminated and erroneous reads in the low coverage data, this suggests that the Kennewick Man's actual PCA coordinates would be closer to the contemporary Northern Native Americans than observed. To provide a statistical argument, we also used a block bootstrap approach. We generated 100

bootstrap data sets using blocks of length 5 MB. For each data set we reproduced the PCA analysis using bammds[44]. In 100/100 cases the closets individual in the PCA analysis was Native American (4 Colville, 5 Cree, and 91 Ojibwa). This illustrates that despite the low coverage, we can assign the sample to the Native American group with strong statistical confidence. However, based on this particular analysis we cannot assign to a single tribe.

7.3 ADMIXTURE analysis

We performed model-based clustering analysis using the maximum-likelihood approach implemented in ADMIXTURE[26]. We ran ADMIXTURE on the full worldwide reference panel of 1,107 individuals, assuming K=2 to K=15 ancestral clusters, and selected the best of 10 replicate runs for each value of K. The ancient individuals were then projected onto the ancestral cluster allele frequencies inferred from the modern individuals as previously described[45]. For the low-coverage Kennewick Man data, we excluded sites were the observed allele corresponded to a damage allele at C→T and G→A SNPs. We again observe only a negligible effect of the lower coverage (Figure S5b) on the estimated cluster proportions. We note that both low coverage individuals share a small proportion of an ancestry cluster related to African populations, which is absent in the high coverage Anzick-1 results. This is likely due to lower quality of the low coverage genotypes, which preferentially contribute ancestry to the most diverged ancestral component.

7.4 Outgroup f_3 - and D-statistics

We used outgroup f_3 - and D-statistics to investigate patterns of admixture and shared ancestry[46]. Standard errors for all statistics were obtained from a block jackknife with 5Mb block size. Modern populations with less than 5 individuals were excluded for these analyses.

8. Estimation of divergence and test of direct ancestry

Rasmus Nielsen and Thorfinn Sand Korneliussen

To test for direct ancestry, we used the method from Rasmussen et al[3]. In brief, this method uses a coalescence model to estimate the branch length of the population tree for two populations, each represented by a single diploid individual. The method makes no assumptions regarding population sizes or other demographic parameters. However, it assumes that neither of the two populations admixed with other populations after they split off from each other. Violations of this assumption would lead to overestimation of branch lengths and increased probability of falsely rejecting the hypothesis of one population being ancestral to the other, when true. The methods achieves its robustness to assumptions regarding demographic history by fully parameterizing the ancestral site frequency spectrum using a free parameter for each possible observation, with the only constraint being that probabilities have to sum to one. The other parameters of the model are the coalescence probabilities along each of the two diverging branches (c_1 and c_2). The hypothesis of a direct ancestry between population 1 and 2 then corresponds parametrically to H_0: $c_1 = 0$, i.e. no coalescences along branch 1 corresponding to a zero length branch length A test of this hypothesis can be conducted using a likelihood ratio test comparing the likelihood under H_0 to the more general hypothesis in which $c_1 \in [0, 1]$.

In order to apply this test, we selected the two Colville reference individuals who show no sign of European admixture (Colville 2 and 8). We then estimated the joint allele frequencies in Kennewick and these individuals, separately for each Colville individual. We did this using a likelihood method described in Nielsen et al.[47] and implemented in the software ANGSD[48]. In brief, the method uses quality scores to provide a maximum likelihood estimate of the joint allele frequencies. As these analyses can be sensitive to un-modeled errors, we removed all transitions. We also trimmed the first and last 3 bp from each read, and filtered using a base quality of 20 and mapping quality of 30 after applying BAQ[49] and adjusting for excessive mismatches (--adjust-MQ 50)[7]. The sequencing data was then used to calculate genotype likelihoods following the samtools[7] model. We then applied the method in Rasmussen et al.[3] to estimate parameters and to obtain likelihood values under the null hypotheses and under the alternative hypothesis (Main text Table 1, Figure S6). For Colville 2 and 8, two times the log likelihood ratios are $\lambda_1 = 19.41$ and $\lambda_2 = 3.93$, respectively. These values should be compared to a 50:50 mixture of a chi-square distribution with one degree of freedom and a point-mass at zero. We can, therefore, reject direct ancestry without any additional gene-flow with strong significance for Colville 2 ($p << 0.01$) and with weak significance for Colville 8 ($p < 0.05$). These calculations assume independence among SNPs. The total number of variable sites, after filtering, used in these analyses is between 87k and 90k for each pair of individuals, giving an average distance between SNPs of approximately 30k bp. There can therefore be some degree of linkage disequilibrium between SNPs that will violate the assumption of independence. If so, the p-values provided here are underestimates of the true p-values. We also notice that the parameter estimates we obtain are quite close to zero. For the second individual the amount of divergence on the Kennewick lineage is quite small. For example, if we assume both populations have the same population size, and that the age of Kennewick Man is 8,500 years, the estimate of the amount of independent divergence in the Kennewick lineage since the split from the Colville population is 691 years.

Our results are compatible with two possible hypotheses:

(1) The Colville individuals are direct descendants of the population to which the Kennewick Man belonged, but have received some relatively minor gene-flow from other Native American populations within the last 8,500 years.

(2) The Colville individuals descend from a population that 8,500 years ago was slightly diverged from the population to which the Kennewick Man belonged.

(3) Of course, it is also possible that there was both some limited divergence and some limited gene-flow which contributed to our results.

To provide a context for the results from the Colville individuals, we also provide estimates to 4 other Native Americans, two Northern Athabascan individuals from Canada[50] and two Karitiana individuals from Brazil[3,4]. In all cases, the estimate of c_1 is substantially larger than for the Colville individuals. There is less evidence against the hypothesis of direct ancestry for Colville individuals than for these other reference individuals. The results of this analysis are shown in main text Table 1.

9. Comparative craniometric analysis of Kennewick Man and inference of population affiliations

Marcia S. Ponce de León and Christoph P. E. Zollikofer

Previous studies of the population affiliations of Kennewick Man have used a wide variety of morphometric and statistical techniques, and a wide range of comparative data from extant and archaeological populations. Based on these studies Kennewick Man has been variably associated with Europeans, Polynesians, and Jomon/Ainu. Overall, there is general agreement that this individual differs substantially from modern Amerind populations, and that it is phenetically close to other proposed Paleoamerican populations[51-55].

These studies are based on two assumptions: (a) The pattern of present-day cranial diversity is an indicator of population structure and history. Accordingly, phenetic similarity of archaeological specimens with present-day groups is potentially informative about their phylogenetic relationships. (b) The Kennewick individual is a typical representative of the population to which it belonged; in other words, its cranial morphology is fairly close to the average morphology of its population and thus provides reliable information about the relationships of the Kennewick population with other populations, both historical and modern.

The first assumption can be tested with independent data such as genetic markers. The second assumption cannot be tested explicitly, but for large samples with known population affiliations it is possible to assess how reliably the affiliation of a given specimen can be reconstructed from its craniometric data. This approach was proposed by W.W. Howells, and tested with specimens drawn at random from his worldwide craniometric data base[56]. The “Kennewick scenario” studied here differs from the "Howells scenario" in that Kennewick is a single representative of an unknown archaeological population, who is compared with a large sample of known-population individuals [in our case the Howells data set[57-59]]. To assess the reliability of population inferences for Kennewick, we thus treat each Howells individual as a "virtual" archaeological specimen with unknown population affiliations, analyze its craniometric relationships with known-population individuals, and compare its inferred affiliation with its known affiliation.

Materials and Methods

The comparative sample consists of the male subset of Howells' craniometric data set (N=1368). Cranial morphology was quantified with the M=35 craniometric variables that could be measured reliably on the Kennewick cranium[55]. Variables are GOL, NOL, BNL, BBH, XCB, XFB, ZYB, AUB, MDH, MDB, FRC, FRS, FRF, PAC, PAS, OCC, OCS, FOL, BPL, NPH, NLH, JUB, NLB, MAB, OBH, OBB, DKB, WNB, ZMB, FMB, NAS, EKB, DKS, IML, XML [see [57] for explanation of variable acronyms]. Principal Components Analysis (PCA) was performed on the covariance matrix of these data (using the covariance rather than the correlation matrix conserves the original metric dimensions of all craniometric variables). PCA shows that the first 12 PCs comprise >95% of the total variation in the sample; higher-order PCs are assumed to represent sampling noise. Craniometric distances between specimens were thus evaluated as Euclidean distances along the first 12 dimensions of PC space.

Analysis 1

We test the hypothesis that Kennewick differs substantially from modern Amerind crania and that it exhibits close phenetic affiliations with Polynesian and Ainu crania[55]. The hypothesis is operationalized in the following way:

Three regional subsets are selected from the Howells data set (males only): Americans [Am] (populations: Arikara, Santa Cruz, Peru), Polynesians [Po] (Mokapu, Easter Island, Moriori, Maori), and Ainu [Ai]. The respective subsample sizes are N_{Ai}=148, N_{Po}=177, and N_{Ai}=48.

Craniometric distances between Kennewick and all American crania of the Howells data set are evaluated, and the respective distance frequency distribution (Ke-[Am]) is evaluated. Similarly, distances between Kennewick and all Polynesian crania of the Howells data set are evaluated, resulting in the distance frequency distribution Ke-[Po]. The same is done for the Ainu, resulting in frequency distribution Ke-[Ai].

Each single American cranium Am_i is treated as a "virtual archaeological specimen" with unknown affiliation. For each Am_i, craniometric distances to all other American specimens are evaluated, resulting in N_{Am} distance frequency distributions Am_i-[Am]. Similarly, for each Am_i, distances to all Polynesian specimens are evaluated, resulting in N_P distance frequency distributions Am_i-[Po], and distances to all Ainu specimens, resulting in Am_i-[Ai].

As shown in figure S7, craniometric distances between Kennewick and Americans (distribution Ke-[Am]) tend to be greater than distances between Kennewick and Polynesians (distribution Ke-[Po]), and between Kennewick and Ainu (distribution Ke-[Ai]); the respective median values are 36.57, 25.16, and 26.98. This confirms earlier observations that Kennewick is more similar to circumpacific than modern Amerind populations[55]. However, since Kennewick is an isolated specimen, the question is whether this pattern of similarity can also be observed in individual modern Amerind specimens. In other words, it remains to be assessed whether Kennewick's distance distributions Ke-[Am], Ke-[Po] and Ke-[Ai] are clearly different from the respective modern Amerind ensembles Am_i-[Am], Am_i-[Po], and Am_i-[Ai]. Fig. S7 shows that Ke-[Am], Ke-[Po] and Ke-[Ai] are enclosed in the 99th-percentile hulls of ensembles Am_i-[Am] and Am_i-[Po] and Am_i-[Ai], respectively. Craniometric data thus do not support the hypothesis that Kennewick Man is an outlier compared to modern Amerinds; rather, he forms part of male Amerind craniometric variation. Note that this does not contradict the well-established observation that the craniometric *mean* of Paleoamerican populations differs from that of modern Amerind populations.

Analysis 2

We assess how reliably population and regional affiliations of single specimens of the Howells sample can be reconstructed with craniometric data. In practical terms, the analysis works as follows: (1) For each individual i (i=1…N) the five craniometric nearest-neighbor individuals (nn1-nn5) are evaluated (excluding as potential nearest neighbors the population subsample to which i belongs). (2) For each nn, population affiliation and regional affiliation are annotated. The same procedures are applied to Kennewick. (3) Similarly, for each individual i (i=1…N), the five nearest-neighbor population-means are evaluated, and their regional affiliation is annotated. The same procedures are applied to Kennewick. (4) The number of cases is recorded for which

the regional affiliation of individual *i* is inferred correctly (*i.e.*, the nearest craniometric neighbor comes from the same region as individual *i*).

Results are summarized in Tables S5-8. Table S5 shows that regional population affiliations of single individuals cannot be inferred with any certainty from the available craniometric data. The probability of correct inference is between 24% (for individual-to-individual comparisons) and 27% (for individual-to-population mean comparisons). For randomized versions of the Howells data set (obtained by random permutation of each individual's population affiliations), the probability is 18%. Our analyses thus confirm the results of earlier studies concerning the inference of population affiliations of single specimens[56,60].

Tables S6 to S8 provide a detailed look at population and regional affiliations of Kennewick and one specific population, the Arikara from North Dakota, who represent the Howells subsample that is geographically closest to Kennewick. Table S6 summarizes the results, while Tables S7 and S8 provide data for each individual Arikara cranium. Kennewick Man shows closest phenetic links with Polynesian populations (again confirming the results of earlier analyses[55]) while the only non-Polynesian individual among his five nearest-neighbors is an Arikara. Results for the male Arikara sample show that the probability for correct regional affiliation of single Arikara individuals is as low as in the global sample (Table S6). Interestingly, the two most frequent nearest-neighbor regional affiliations of the Arikara (America and Polynesia), have about equal probabilities, and various Arikara individuals (e.g. nrs. 770, 775 in Table S7) exhibit a pattern of Polynesian affiliation that is similar to Kennewick.

Overall, our results indicate that population affiliations of single crania cannot be inferred reliably from the set of craniometric variables established by W. W. Howells. These findings apply to representatives of Paleoamerican populations such as Kennewick, and also to representatives of modern Amerind and other worldwide populations. In other words, if the Arikara population would be known from only a single cranium, craniometric inferences on its closest living relatives would be as equivocal as for Kennewick. The observation that Kennewick exhibits close craniometric affiliations with Polynesians and Ainu, and that equally close links can be found between some Arikara individuals and Polynesians/Ainu, requires further consideration. However, such similarities are unlikely to reflect close common ancestry of Amerind populations and Polynesians/Ainu. They more likely reflect phenotypic similarity as an effect of mutation and drift, or as an effect of similar processes of adaptation and/or *in-vivo* modification.

10. Sample collection and community engagement

Thomas W. Stafford, Jr. and Eske Willerslev

As part of the scientific study of the Kennewick remains, TS obtained samples for AMS ^{14}C, stable isotope and aDNA analysis; the radiocarbon and isotope results were published in 2014[61,62]. Based on these results and additional examination of the skeletal remains by TS (Owsley, et al.[63]) TS selected the first of two samples for aDNA analysis by EW and CV.

When this initial sample indicated DNA was present, TS visited the Burke Museum June 23-24, 2013 and selected a second bone for aDNA testing. Based on physical examinations and previous chemical analyses, TS chose the illustrated sample, which the COE subsequently sent to EW for aDNA analysis.

The specimen selected appeared the most suitable for aDNA and was bone that had previously been sampled for aDNA by Yale University. Therefore, no new bone was destroyed during the present aDNA work. The genome reported here is derived from a small piece of the specimen shown in Figure S8, the proximal one-half of a left, third metacarpal having specimen number SR-7091 (TS nomenclature) and COE curation number 97.L.16(Mca). As it is not possible for us to receive more sample for DNA analyses, the extracted DNA was sequenced to saturation. This provided us with approximately 1X coverage of the genome.

When the initial DNA results from that analysis indicated that Kennewick Man was more closely related to Native Americans than to other worldwide populations, EW approached members of the Claimant Plateau tribes (who claim ancestry and had requested repatriation under the Native American Graves Protection and Repatriation Act [NAGPRA]). He met in the US with their representatives to report and discuss the preliminary findings. Subsequently, members of the Claimant Plateau tribes visited the Centre for GeoGenetics in Copenhagen. Since then, EW has been in regular contact with them, and particularly the Confederated Tribes of the Colville Reservation (Colville) who provided DNA for ancestry comparison to the genome data of the Kennewick Man. The Colville DNA collection was done by the members of the tribe. The researchers do not know the identity of the two dozen individuals who provided the samples. The agreement between EW and the Colville tribe is that the SNP chip data of the tribal members can be made available for other researchers, but only for confirmation of the claimed ancestry with the Ancient One.

References

1. Rohland, N. & Hofreiter, M. Ancient DNA extraction from bones and teeth. *Nat Protoc* **2,** 1756–1762 (2007).
2. Orlando, L. *et al.* Revising the recent evolutionary history of equids using ancient DNA. *Proc Natl Acad Sci U S A* **106,** 21754–21759 (2009).
3. Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506,** 225–229 (2014).
4. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* (2012). doi:10.1126/science.1224344
5. Lindgreen, S. AdapterRemoval: Easy Cleaning of Next Generation Sequencing Reads. *BMC Res Notes* **5,** 337 (2012).
6. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
7. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
8. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci USA* **104,** 14616–14621 (2007).
9. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. & Orlando, L.

mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* (2013). doi:10.1093/bioinformatics/btt193
10. Kloss-Brandstätter, A. *et al.* HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum Mutat* **32,** 25–32 (2011).
11. Bandelt, H. J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16,** 37–48 (1999).
12. Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23,** 147 (1999).
13. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* **30,** E386–94 (2009).
14. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr Biol* **23,** 553–559 (2013).
15. Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc R Soc B* **279,** 4724–4733 (2012).
16. Deagle, B. E., Eveson, J. P. & Jarman, S. N. Quantification of damage in DNA recovered from highly degraded samples--a case study on DNA in faeces. *Front Zool* **3,** 11 (2006).
17. Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507,** 225–228 (2014).
18. Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499,** 74–78 (2013).
19. Pedersen, J. S. *et al.* Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res* **24,** 454–466 (2014).
20. Schroeder, H. *et al.* Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean. *Proc Natl Acad Sci U S A* (2015). doi:10.1073/pnas.1421784112
21. Lindahl, T. & Andersson, A. Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. *Biochemistry* **11,** 3618–3623 (1972).
22. Lindahl, T. & Nyberg, B. Rate of depurination of native deoxyribonucleic acid. *Biochemistry* **11,** 3610–3618 (1972).
23. Smith, C. I., Chamberlain, A. T., Riley, M. S., Stringer, C. & Collins, M. J. The thermal history of human fossils and the likelihood of successful DNA amplification. *Journal of Human Evolution* **45,** 203–217 (2003).
24. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463,** 757–762 (2010).
25. Busing, F. M., Meijer, E. & Leeden, R. V. D. Delete-m Jackknife for Unequal m. *Statistics and Computing* **9,** 3–8 (1999).
26. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19,** 1655–1664 (2009).
27. Zegura, S. L., Karafet, T. M., Zhivotovsky, L. A. & Hammer, M. F. High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Mol Biol Evol* **21,** 164–175 (2004).
28. Bolnick, D. A., Bolnick, D. I. & Smith, D. G. Asymmetric male and female

genetic histories among Native Americans from Eastern North America. *Mol Biol Evol* **23,** 2161–2174 (2006).
29. Cann, H. M. *et al.* A human genome diversity cell line panel. *Science* **296,** 261–262 (2002).
30. Poznik, G. D. *et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341,** 562–565 (2013).
31. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
32. Dulik, M. C. *et al.* Y-chromosome analysis reveals genetic divergence and new founding native lineages in Athapaskan- and Eskimoan-speaking populations. *Proc Natl Acad Sci U S A* **109,** 8471–8476 (2012).
33. Regueiro, M., Alvarez, J., Rowold, D. & Herrera, R. J. On the origins, rapid expansion and genetic diversity of Native Americans from hunting-gatherers to agriculturalists. *Am J Phys Anthropol* **150,** 333–348 (2013).
34. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488,** 370–374 (2012).
35. Verdu, P. *et al.* Patterns of admixture and population structure in native populations of northwest north america. *PLoS Genet* **10,** e1004530 (2014).
36. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
37. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10,** 5–6 (2013).
38. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am J Hum Genet* **93,** 278–288 (2013).
39. Japanese Archipelago Human Population Genetics Consortium *et al.* The history of human populations in the Japanese Archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. *J Hum Genet* **57,** 787–795 (2012).
40. Wollstein, A. *et al.* Demographic history of Oceania inferred from genome-wide data. *Curr Biol* **20,** 1983–1992 (2010).
41. Kayser, M. *et al.* Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *Am J Hum Genet* **82,** 194–198 (2008).
42. Fedorova, S. A. *et al.* Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *BMC Evol Biol* **13,** 127 (2013).
43. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2,** e190 (2006).
44. Malaspinas, A.-S. *et al.* bammds: A tool for assessing the ancestry of low depth whole genome data using multidimensional scaling (MDS). *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu410
45. Sikora, M. *et al.* Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the tyrolean iceman and the genetic structure of europe. *PLoS Genet* **10,** e1004353 (2014).
46. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192,** 1065–1093 (2012).

47. Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y. & Wang, J. SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. *PLoS ONE* **7,** e37558 (2012).
48. Korneliussen, T., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15,** 356 (2014).
49. Li, H. Improving SNP discovery by base alignment quality. *Bioinformatics* **27,** 1157–1158 (2011).
50. Raghavan, M. *et al.* The genetic prehistory of the New World Arctic. *Science* **345,** 1255832 (2014).
51. Brace, C. L. *et al.* in *Kennewick Man* (Owsley, D. W. & Jantz, R. L.) 461–471 (Texas A&M University Press, 2014).
52. Chatters, J. C. The recovery and first analysis of an Early Holocene human skeleton from Kennewick, Washington. *Am Antiq* **65,** 291–316 (2000).
53. Gill, G. W. in *Kennewick Man* (Owsley, D. W. & Jantz, R. L.) 503–518 (Texas A&M University Press, 2014).
54. Jantz, R. L. & Owsley, D. W. Reply to Van Vark et al.: Is European Upper Paleolithic cranial morphology a useful analogy for early Americans? *Am J Phys Anthropol* **121,** 185–188 (2003).
55. Jantz, R. L. & Spradley, M. K. in *Kennewick Man* (Owsley, D. W. & Jantz, R. L.) 472–491 (Texas A&M University Press, 2014).
56. Howells, W. W. *Who's Who in Skulls*. (Peabody Museum of Archaeology &, 1995).
57. Howells, W. W. *Cranial Variation in Man*. (Peabody Museum of Archaeology and Ethnology, Harvard University Publications Department, 1973).
58. Howells, W. W. *Skull Shapes and the Map*. (Peabody Museum of Archaeology &, 1989).
59. Howells, W. W. Howells' craniometric data on the Internet. *Am J Phys Anthropol* **101,** 441–442 (1996).
60. van Vark, G. N., Kuizenga, D. & Williams, F. L. Kennewick and Luzia: lessons from the European Upper Paleolithic. *Am J Phys Anthropol* **121,** 181–4– discussion 185–8 (2003).
61. Stafford, T. W., Jr. in *Kennewick Man* (Owsley, D. W. & Jantz, R. L.) 59–89 (Texas A&M University Press, 2014).
62. Schwarcz, H. P. *et al.* in *Kennewick Man* (Owsley, D. W. & Jantz, R. L.) 310–322 (Texas A&M University Press, 2014).
63. Owsley, D. W. & Stafford, T. W., Jr. in *Kennewick Man* (Owsley, D. W. & Jantz, R. L.) 323–381 (Texas A&M University Press, 2014).

Table S1. Sequencing and mapping summary.

Library name	Raw reads	Trimmed reads longer than 30bp	Mapped human Q30	Duplicates removed
H37S_Ken27_DSL	152,049,054	137,356,489	615,220	537,870
H37S_Ken31_DSL	162,635,384	147,527,105	687,122	586,506
S7QJ_Ken19_SSL	2,644,328,735	1,976,296,206	28,594,631	27,118,425
S7QJ_Ken26_SSL	3,061,525,209	2,291,280,205	33,501,220	31,756,293
Total	6,020,538,382	4,552,460,005	63,398,193	59,999,094

Table S2: DNA decay rates

Comparison of DNA decay in genomic data from four different skeletons. Approximate ages and estimated annual surface temperatures for the sites are listed. Lambda (λ) is the DNA damage fraction (per site) estimated from the sequence length distributions (Figure S2), and the estimated average DNA fragment length in the sample is calculated as 1/λ. Lambda is converted to decay rate (*k*, per site per year) by dividing with sample age. Molecular half-lives for 100 bp fragments are calculated as in Allentoft et al[15]. La Braña results are based on data from Olalde et al.[17], Anzick results from Rasmussen et al.[3], and Caribbean from Schroeder et al[20].

	App. age, yrs	App. temp.	λ	Av. length	k	k, 100 bp	Half-life (yrs), 100 bp
Caribbean	340	27.0°C	0,014	71 bp	4,12E-05	4,11E-03	169
La Brana, Spain	7500	8.1°C	0,033	30 bp	4,40E-06	4,40E-04	1576
Kennewick, Washington	9000	12.5°C	0,017	59 bp	1,89E-06	1,89E-04	3670
Anzick, Montana	12785	4.8°C	0,018	56 bp	1,41E-06	1,41E-04	4916
Kennewick, SSL	9000	12.5°C	0,044	23 bp	4,89E-06	4,89E-04	1418
Kennewick, SSL mtDNA	9000	12.5°C	0,032	31 bp	3,56E-06	3,55E-04	1950

Table S3: Summary of major and minor read counts in the known polymorphic sites included in the analysis and their adjacent sites.

	Position relative to known polymorphic sites								
	-4	-3	-2	-1	0	1	2	3	4
Minor reads	37	40	38	52	98	51	37	40	42
Major reads	6803	6808	6822	6811	6760	6821	6809	6796	6791

Table S4. Estimated contamination rates and posterior probabilities from the five 1000 genomes population being the contamination source.

Population *y*	Estimated contamination rate, θ assuming *y* is the source	Probability of source population, P(pop=*y*|*θ, ε,f*)
YRI	0.018	0.0026
PEL	0.028	0.0107
CHB	0.024	0.0012
CEU	0.025	0.9565
GIH	0.024	0.0291

Table S5. Nearest-neighbor relationships among male Howells individuals

nearest neighbor of a given individual is:	***n***	**ratio (*n*/1368)***
an individual of a population from the same region	310	0.23
the mean of a population from the same region	376	0.27
an individual with randomly assigned region	251	0.18*

* mean value of resampling the Howells data set 100 times, with random permutations of population affiliations

Table S6. Nearest-neighbor relationships of male Arikara individuals

nearest neighbor of male Arikara individual is:	*n*	ratio (*n*/42)*
an individual from an American population	10	0.24
an individual from a Polynesian population	9	0.21
the population-mean of an American population	10	0.24
the population-mean of a Polynesian population	11	0.26

Table S7. Craniometric affinities of Arikara males and Kennewick with worldwide individuals

Howells id	population	cranial size	neighbor1	neighbor2	neighbor3	neighbor4	neighbor5	region1	region2	region3	region4	region5	Am = region1	Am in region1-5	Polynesia in region1-5
762	Arikara	47.27	Egypt	SJapan	Philipp	S Cruz	Berg	NAf+Eur	Asia	Asia	Am	NAf+Eur	0	1	0
763	Arikara	49.24	Moriori	Peru	Peru	S Cruz	NJap	Polynesia	Am	Am	Am	Asia	0	3	1
764	Arikara	49.40	S Cruz	Eskimo	Teita	Australi	Atayal	Am	Arctic	S-Afr	Sahul	Asia	1	1	0
765	Arikara	48.43	Hainan	Atayal	Berg	Eskimo	Atayal	Asia	Asia	NAf+Eur	Arctic	Asia	0	0	0
766	Arikara	50.68	Zulu	Anyang	Tolai	Guam	Hainan	S-Afr	Asia	Sahul	Asia	Asia	0	0	0
767	Arikara	49.34	Teita	Moriori	Moriori	Philipp	S Cruz	S-Afr	Polynesia	Polynesia	Asia	Am	0	1	2
768	Arikara	49.61	Hainan	Berg	Dogon	Zulu	Peru	Asia	NAf+Eur	S-Afr	S-Afr	Am	0	1	0
769	Arikara	47.47	Philipp	Peru	Norse	Eskimo	S Cruz	Asia	Am	NAf+Eur	Arctic	Am	0	2	0
770	Arikara	51.36	Moriori	Moriori	Moriori	Moriori	Moriori	Polynesia	Polynesia	Polynesia	Polynesia	Polynesia	0	0	5
771	Arikara	50.04	S Cruz	Moriori	Buriat	Guam	Moriori	Am	Polynesia	Arctic	Asia	Polynesia	1	1	2
772	Arikara	49.16	Buriat	Buriat	Buriat	Guam	Buriat	Arctic	Arctic	Arctic	Asia	Arctic	0	0	0
773	Arikara	49.91	Anyang	Tasman	Philipp	Hainan	Tolai	Asia	Sahul	Asia	Asia	Sahul	0	0	0
774	Arikara	48.91	Moriori	Tolai	S Cruz	S Cruz	S Cruz	Polynesia	Sahul	Am	Am	Am	0	3	1
775	Arikara	52.47	Moriori	Moriori	SMaori	Moriori	SJapan	Polynesia	Polynesia	Polynesia	Polynesia	Asia	0	0	4
776	Arikara	50.66	Buriat	Anyang	Peru	Guam	NJap	Arctic	Asia	Am	Asia	Asia	0	1	0
777	Arikara	48.60	SJapan	Philipp	Moriori	Moriori	Hainan	Asia	Asia	Polynesia	Polynesia	Asia	0	0	2
778	Arikara	50.31	Moriori	Moriori	NJap	Peru	Eskimo	Polynesia	Polynesia	Asia	Am	Arctic	0	1	2
779	Arikara	48.23	Peru	Peru	Andam	Australi	Moriori	Am	Am	Asia	Sahul	Polynesia	1	2	1
780	Arikara	49.11	S Cruz	Norse	Peru	Philipp	Anyang	Am	NAf+Eur	Am	Asia	Asia	1	2	0
781	Arikara	48.83	Peru	Hainan	Atayal	NJap	Hainan	Am	Asia	Asia	Asia	Asia	1	1	0
782	Arikara	48.82	Peru	Moriori	SJapan	Ainu	Peru	Am	Polynesia	Asia	Asia	Am	1	2	1
783	Arikara	49.65	SJapan	Anyang	Zalavar	Egypt	Peru	Asia	Asia	NAf+Eur	NAf+Eur	Am	0	1	0
784	Arikara	48.74	Norse	Andam	Tolai	SJapan	Mokap	NAf+Eur	Asia	Sahul	Asia	Polynesia	0	0	1
785	Arikara	49.97	Australi	Australi	Australi	Tolai	Guam	Sahul	Sahul	Sahul	Sahul	Asia	0	0	0
786	Arikara	48.29	Peru	Moriori	Norse	Norse	NJap	Am	Polynesia	NAf+Eur	NAf+Eur	Asia	1	1	1
787	Arikara	50.11	Anyang	S Cruz	Hainan	Peru	Anyang	Asia	Am	Asia	Am	Asia	0	2	0
788	Arikara	51.20	Moriori	Mokap	Buriat	Ainu	Anyang	Polynesia	Polynesia	Arctic	Asia	Asia	0	0	2
789	Arikara	47.22	Philipp	Tasman	Peru	Philipp	Andam	Asia	Sahul	Am	Asia	Asia	0	1	0
790	Arikara	48.07	Hainan	Philipp	Hainan	Zalavar	Peru	Asia	Asia	Asia	NAf+Eur	Am	0	1	0
791	Arikara	52.23	SMaori	Ainu	Ainu	SMaori	Hainan	Polynesia	Asia	Asia	Polynesia	Asia	0	0	2
792	Arikara	48.84	Anyang	Philipp	S Cruz	S Cruz	Guam	Asia	Asia	Am	Am	Asia	0	2	0
793	Arikara	51.26	Moriori	Hainan	Dogon	Hainan	Peru	Polynesia	Asia	S-Afr	Asia	Am	0	1	1
794	Arikara	49.33	Anyang	S Cruz	S Cruz	Peru	Buriat	Asia	Am	Am	Am	Arctic	0	3	0
795	Arikara	50.19	S Cruz	Moriori	NJap	Norse	SMaori	Am	Polynesia	Asia	NAf+Eur	Polynesia	1	1	2
796	Arikara	47.25	Buriat	Atayal	Berg	Anyang	Hainan	Arctic	Asia	NAf+Eur	Asia	Asia	0	0	0
797	Arikara	47.16	Peru	S Cruz	S Cruz	Peru	S Cruz	Am	Am	Am	Am	Am	1	5	0
798	Arikara	46.47	Anyang	Hainan	Teita	Eskimo	Hainan	Asia	Asia	S-Afr	Arctic	Asia	0	0	0
799	Arikara	50.32	Hainan	Moriori	Eskimo	SMaori	S Cruz	Asia	Polynesia	Arctic	Polynesia	Am	0	1	2
800	Arikara	49.31	Eskimo	NJap	Atayal	SJapan	Ainu	Arctic	Asia	Asia	Asia	Asia	0	0	0
801	Arikara	46.05	Mokap	Peru	S Cruz	Peru	S Cruz	Polynesia	Am	Am	Am	Am	0	4	1
802	Arikara	48.98	Guam	NJap	Peru	Easter I	Buriat	Asia	Asia	Am	Polynesia	Arctic	0	1	1
803	Arikara	50.40	Peru	Berg	Philipp	Peru	Moriori	Am	NAf+Eur	Asia	Am	Polynesia	1	2	1
	Kennewick	49.59	Moriori	Moriori	Easter I	Arikara	Mokap	Polynesia	Polynesia	Polynesia	Am	Polynesia	0	1	4

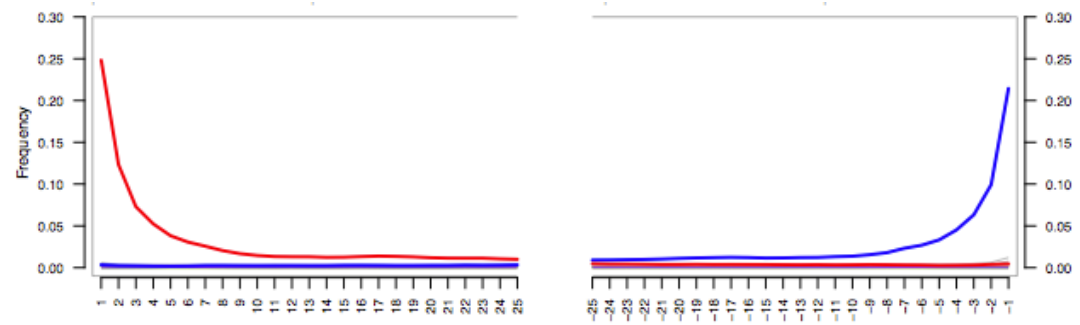
*explanation of variables: Howells id: specimen nr. in Howells' global data set; cranial size: geometric mean of craniometric variables; neighbor1-5: nearest-neighbor populations; region1-5: nearest-neighbor regions; Am=region1: nearest neighbor region is Americas; Am in region1-5: Americas among the five nearest neighbor regions; Polynesia in region1-5: Polynesia among the five nearest neighbor regions

Table S8. Craniometric affinities of Arikara males and Kennewick with worldwide population means.

Howells id	population	cranial size	neighbor1	neighbor2	neighbor3	neighbor4	neighbor5	region1	region2	region3	region4	region5	Am = region1	Am in region1-5	Polynesia in region1-5
762	Arikara	47.27	Zalavar	Moriori	Egypt	Berg	Buriat	NAf+Eur	Polynesia	NAf+Eur	NAf+Eur	Arctic	0	0	1
763	Arikara	49.24	Moriori	Philipp	Andaman	Tolai	Hainan	Polynesia	Asia	Asia	Sahul	Asia	0	0	1
764	Arikara	49.40	Eskimo	SMaori	Peru	Atayal	S Cruz	Arctic	Polynesia	Amer	Asia	Amer	0	2	1
765	Arikara	48.43	Atayal	Peru	Hainan	NJapan	SMaori	Asia	Amer	Asia	Asia	Polynesia	0	1	1
766	Arikara	50.68	Teita	Hainan	Zulu	SMaori	Guam	S-Afr	Asia	S-Afr	Polynesia	Asia	0	0	1
767	Arikara	49.34	Moriori	Buriat	Egypt	Australi	S Cruz	Polynesia	Arctic	NAf+Eur	Sahul	Amer	0	1	1
768	Arikara	49.61	Zulu	Hainan	Berg	Teita	Dogon	S-Afr	Asia	NAf+Eur	S-Afr	S-Afr	0	0	0
769	Arikara	47.47	Hainan	Philipp	Zalavar	Buriat	SJapan	Asia	Asia	NAf+Eur	Arctic	Asia	0	0	0
770	Arikara	51.36	Moriori	Egypt	Easter I	SJapan	Mokap	Polynesia	NAf+Eur	Polynesia	Asia	Polynesia	0	0	3
771	Arikara	50.04	Buriat	Moriori	Mokap	Philipp	Berg	Arctic	Polynesia	Polynesia	Asia	NAf+Eur	0	0	2
772	Arikara	49.16	Buriat	Mokap	Berg	Philipp	Zalavar	Arctic	Polynesia	NAf+Eur	Asia	NAf+Eur	0	0	1
773	Arikara	49.91	Hainan	Moriori	Zalavar	Teita	Philipp	Asia	Polynesia	NAf+Eur	S-Afr	Asia	0	0	1
774	Arikara	48.91	S Cruz	Moriori	NMaori	SJapan	Peru	Amer	Polynesia	Polynesia	Asia	Amer	1	2	2
775	Arikara	52.47	Peru	Moriori	SMaori	S Cruz	Zulu	Amer	Polynesia	Polynesia	Amer	S-Afr	1	2	2
776	Arikara	50.66	Buriat	Mokap	Philipp	Tasma	Hainan	Arctic	Polynesia	Asia	Sahul	Asia	0	0	1
777	Arikara	48.60	SMaori	Hainan	Berg	Dogon	NMaori	Polynesia	Asia	NAf+Eur	S-Afr	Polynesia	0	0	2
778	Arikara	50.31	Hainan	Moriori	Guam	Teita	NMaori	Asia	Polynesia	Asia	S-Afr	Polynesia	0	0	2
779	Arikara	48.23	Moriori	NMaori	Egypt	Zalavar	Hainan	Polynesia	Polynesia	NAf+Eur	NAf+Eur	Asia	0	0	2
780	Arikara	49.11	Moriori	Berg	Zalavar	Buriat	Tasma	Polynesia	NAf+Eur	NAf+Eur	Arctic	Sahul	0	0	1
781	Arikara	48.83	Peru	NJapan	Hainan	Guam	Anyang	Amer	Asia	Asia	Asia	Asia	1	1	0
782	Arikara	48.82	NJapan	Moriori	SJapan	Atayal	Peru	Asia	Polynesia	Asia	Asia	Amer	0	1	1
783	Arikara	49.65	Philipp	Anyang	SJapan	Hainan	NJapan	Asia	Asia	Asia	Asia	Asia	0	0	0
784	Arikara	48.74	Hainan	SMaori	NMaori	Moriori	Guam	Asia	Polynesia	Polynesia	Polynesia	Asia	0	0	3
785	Arikara	49.97	Anyang	Australi	Guam	NMaori	S Cruz	Asia	Sahul	Asia	Polynesia	Amer	0	1	1
786	Arikara	48.29	Moriori	Egypt	Zalavar	Mokap	Easter I	Polynesia	NAf+Eur	NAf+Eur	Polynesia	Polynesia	0	0	3
787	Arikara	50.11	Peru	S Cruz	Moriori	Hainan	Philipp	Amer	Amer	Polynesia	Asia	Asia	1	2	1
788	Arikara	51.20	Buriat	Moriori	Zalavar	Egypt	Berg	Arctic	Polynesia	NAf+Eur	NAf+Eur	NAf+Eur	0	0	1
789	Arikara	47.22	Tolai	Peru	Philipp	Hainan	Andaman	Sahul	Amer	Asia	Asia	Asia	0	1	0
790	Arikara	48.07	Hainan	Moriori	NJapan	Peru	Zalavar	Asia	Polynesia	Asia	Amer	NAf+Eur	0	1	1
791	Arikara	52.23	SMaori	Berg	NMaori	Ainu	Peru	Polynesia	NAf+Eur	Polynesia	Asia	Amer	0	1	2
792	Arikara	48.84	Tasma	Philipp	Mokap	NMaori	Zalavar	Sahul	Asia	Polynesia	Polynesia	NAf+Eur	0	0	2
793	Arikara	51.26	Norse	Peru	Berg	SMaori	Moriori	NAf+Eur	Amer	NAf+Eur	Polynesia	Polynesia	0	1	2
794	Arikara	49.33	Guam	Anyang	Hainan	Buriat	S Cruz	Asia	Asia	Asia	Arctic	Amer	0	1	0
795	Arikara	50.19	SMaori	NMaori	Guam	Moriori	S Cruz	Polynesia	Polynesia	Asia	Polynesia	Amer	0	1	3
796	Arikara	47.25	Eskimo	Hainan	Atayal	NJapan	Peru	Arctic	Asia	Asia	Asia	Amer	0	1	0
797	Arikara	47.16	NJapan	S Cruz	Buriat	SJapan	Peru	Asia	Amer	Arctic	Asia	Amer	0	2	0
798	Arikara	46.47	Eskimo	Hainan	Atayal	SMaori	Moriori	Arctic	Asia	Asia	Polynesia	Polynesia	0	0	2
799	Arikara	50.32	S Cruz	Anyang	NJapan	NMaori	Peru	Amer	Asia	Asia	Polynesia	Amer	1	2	1
800	Arikara	49.31	Atayal	Hainan	NJapan	Eskimo	Anyang	Asia	Asia	Asia	Arctic	Asia	0	0	0
801	Arikara	46.05	SJapan	NJapan	S Cruz	Ainu	Philipp	Asia	Asia	Amer	Asia	Asia	0	1	0
802	Arikara	48.98	Mokap	Philipp	Buriat	Dogon	Easter I	Polynesia	Asia	Arctic	S-Afr	Polynesia	0	0	2
803	Arikara	50.40	Moriori	Berg	Tolai	Norse	Hainan	Polynesia	NAf+Eur	Sahul	NAf+Eur	Asia	0	0	1
mean	Arikara	49.26	Hainan	Moriori	Zalavar	Peru	N Japan	Asia	Polynesia	NAf+Eur	Amer	Asia	0	1	1
-	Kennewick	49.59	Moriori	NJapan	Australi	SJapan	NMaori	Polynesia	Asia	Sahul	Asia	Polynesia	0	0	2

*explanation of variables: Howells id: specimen nr. in Howells' global data set; cranial size: geometric mean of craniometric variables; neighbor1-5: nearest-neighbor populations; region1-5: nearest-neighbor regions; Am=region1: nearest neighbor region is Americas; Am in region1-5: Americas among the five nearest neighbor regions; Polynesia in region1-5: Polynesia among the five nearest neighbor regions

a



b

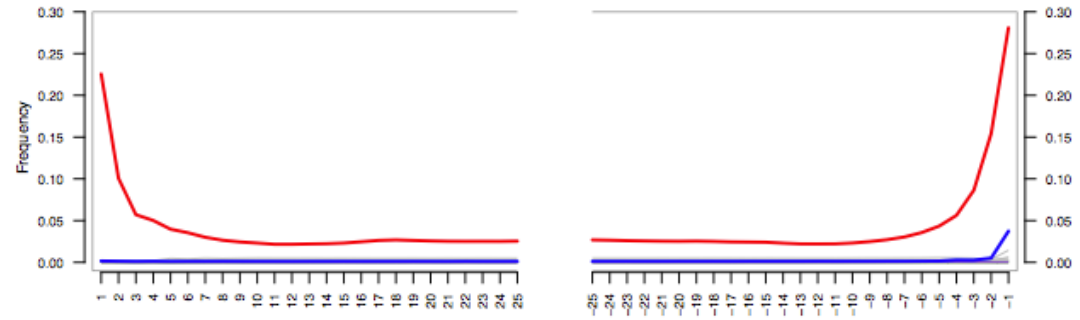


Figure S1. Damage patterns for Kennewick Man. a, for double strand libraries. **b**, for single strand libraries. Mismatch frequency to the reference as function of relative position within the read position, C→T in red and G→A in blue.

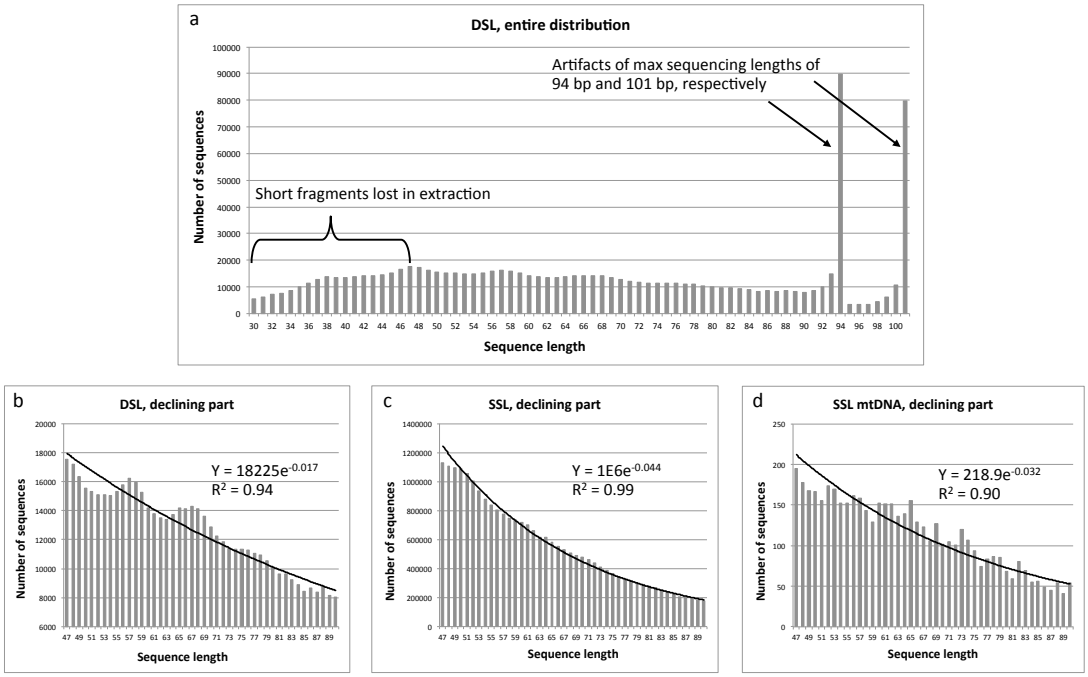


Figure S2. DNA fragment length distributions. Fragment length distributions from the Kennewick Man genomic data. **a**, **b**, represent DSL data. Only the declining part of the distribution (**b**) was used for the decay estimate since the artifacts illustrated in the entire distribution (**a**) must be removed. **c**, **d**, represent SSL data with the distribution depicted for nuclear DNA and mtDNA respectively.

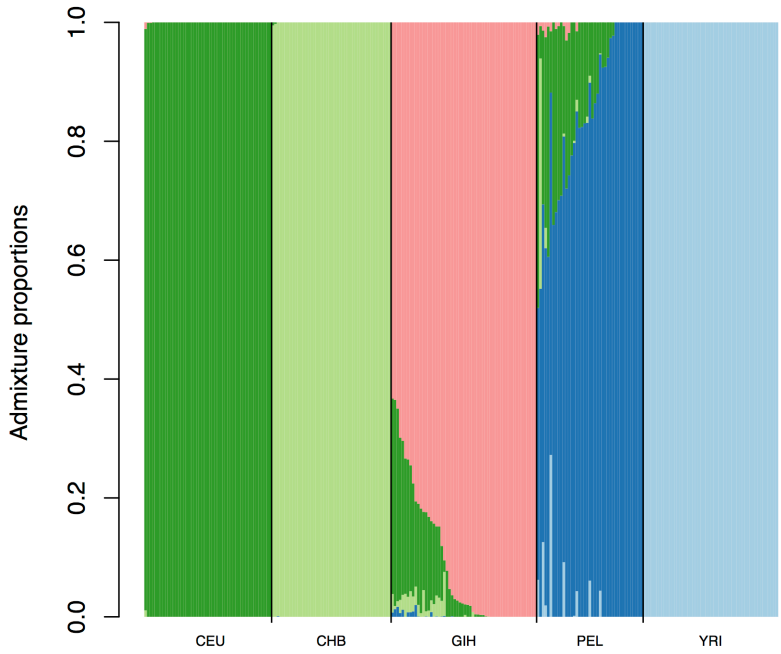


Figure S3. Admixture proportions for the five 1000 genomes project used in the contamination analyses. The admixture proportions were estimated using ADMIXTURE.

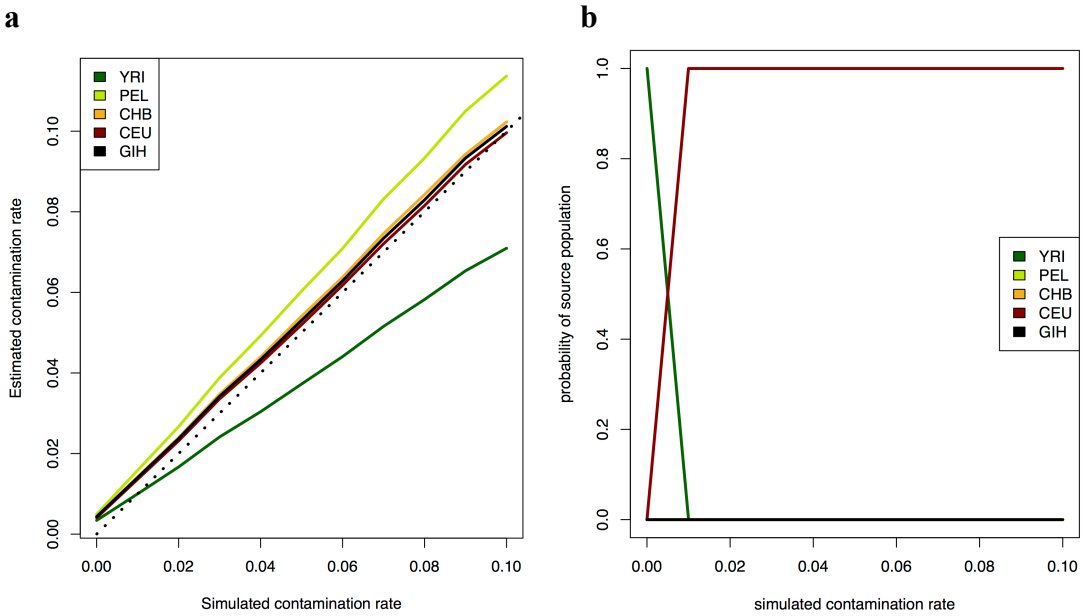


Figure S4. Simulated contamination based on real data. a, The simulated contamination rates plotted against the estimated contamination rates using each of the five 1000 genomes populations in the *Popset* as the assumed contamination source populations. The dashed line is the x=y line and thus shows where the simulated contamination rates equals the estimated contamination rates. **b,** The simulated contamination rates plotted against the estimated posterior probability that each of the five populations in *Popset*.

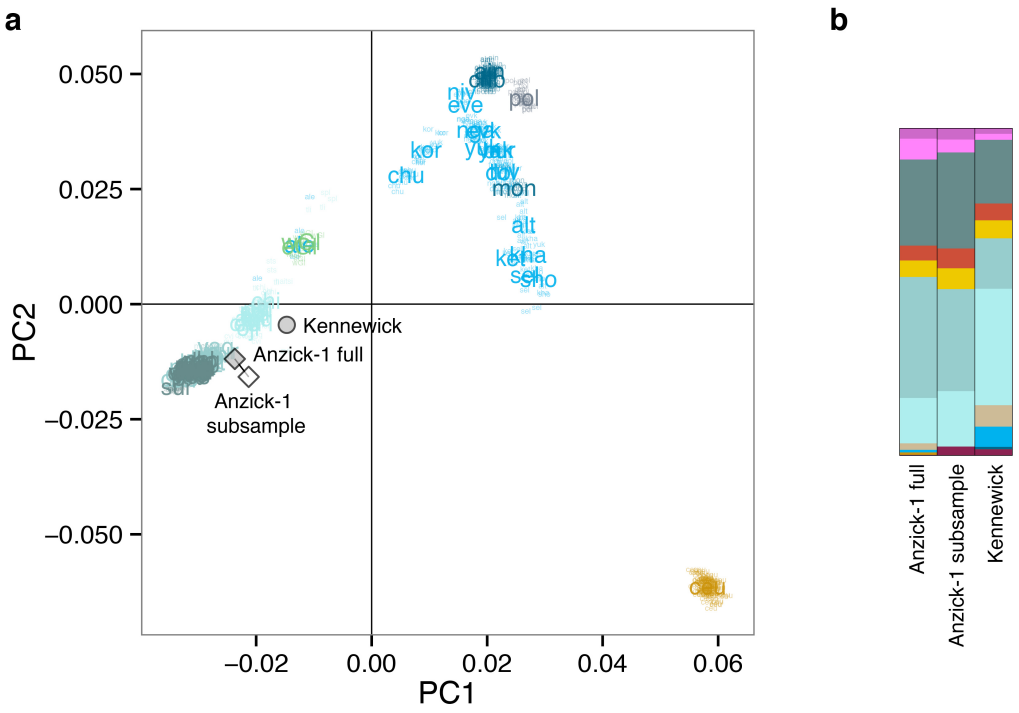


Figure S5. Effect of low coverage on Anzick-1 ancestry. Results of (**a**) PCA and (**b**) ADMIXTURE analyses comparing the subsampled Anzick-1 data (1.6X coverage) to the full data (14.4X coverage).

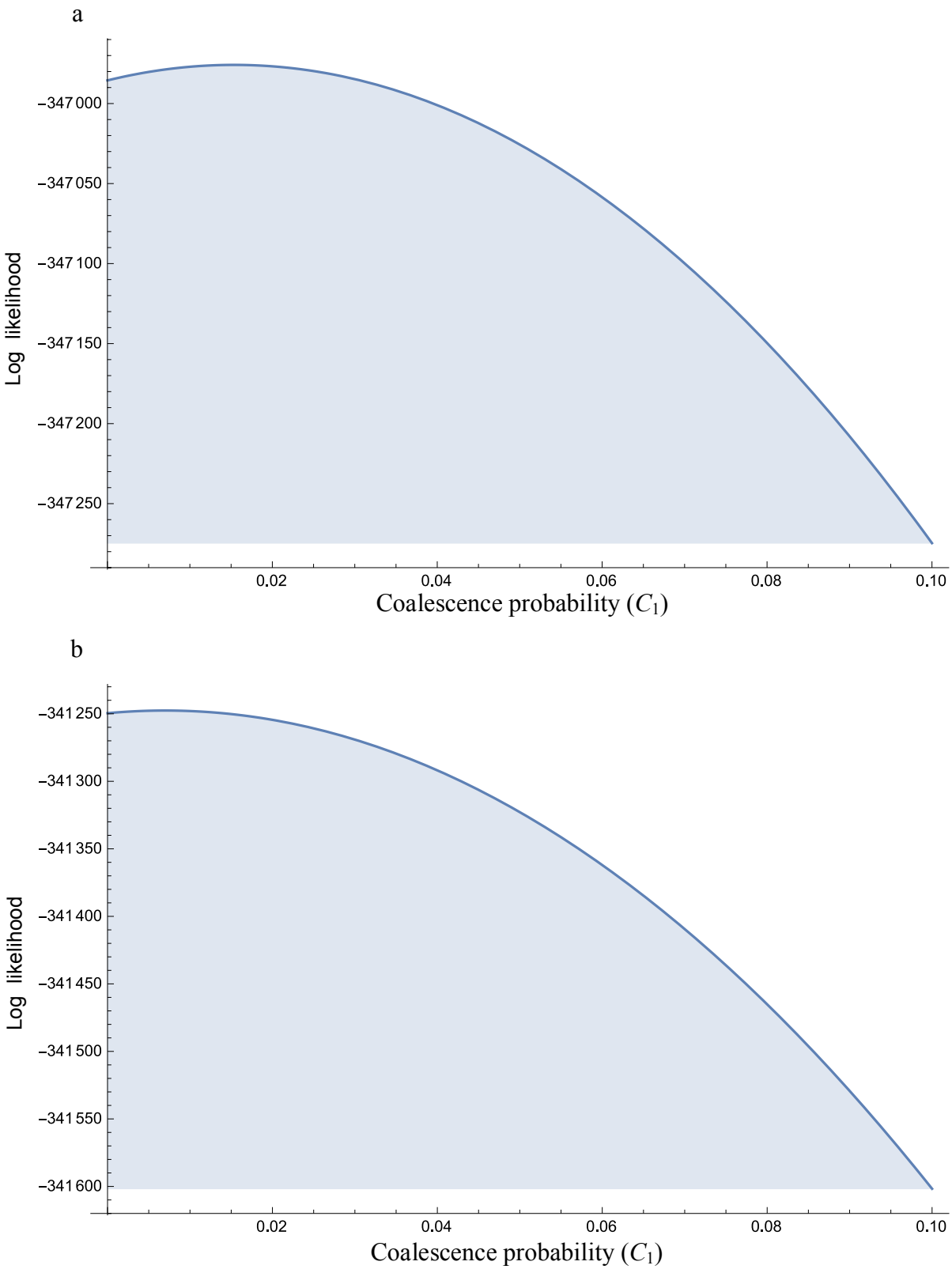


Figure S6. Profile likelihood functions for the coalescence probability in the Kennewick lineage (c_1) relative to Colville 2 and 8, panel a and b, respectively.

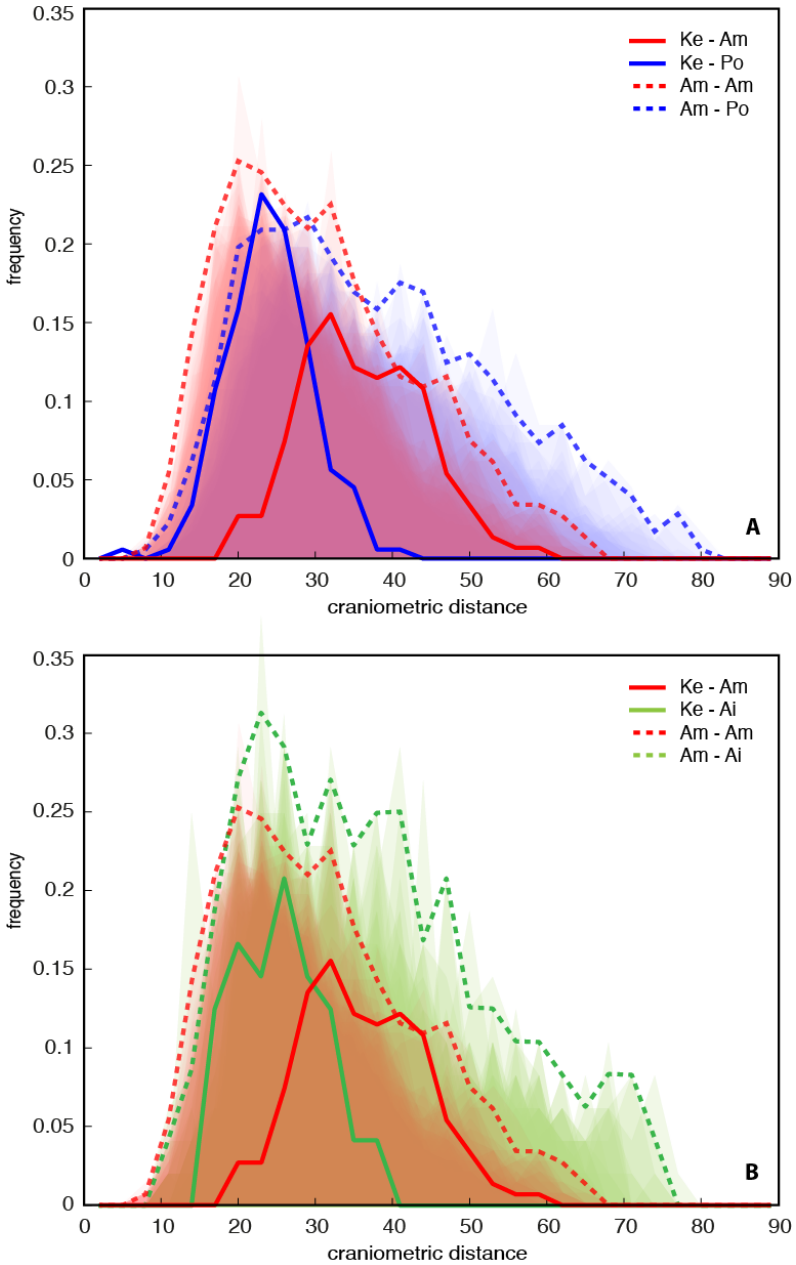


Figure S7. Frequency distributions (FDs) of craniometric distances between Kennewick (Ke), Native American (Am), Polynesian (Po), and Ainu (Ai) specimens. Solid lines: FD of distances from Ke to all Am (red), all Po (blue), and all Ai (green). Colored areas: FDs of distances from individual Am specimens to all other Am (red), all Po (blue), and all Ai (green). Dashed lines: 99th-percentile hulls of the Am-Am (red), the Am-Po (blue), and Am-Ai (green) FD ensembles.

Figure S8. Kennewick Man partial fragment of proximal, left 3rd metacarpal SR-7091, 97.L.16(Mca) used for AMS ^{14}C dating and a small fragment was sent for DNA analyses.